1 Model Validation

- It's useful to ***validate*** (check) a model that's been selected using the criteria of Class Notes 20. Two ways to do this are:

  1. Collect a ***new data set***, independent of the one used to select and fit the model, and then see how well the model fitted to the original data set fits this new one.
  2. ***Cross-validation*** (aka ***data splitting***): Before selecting and fitting a model, randomly split the data set into two parts, a ***training set*** and a ***validation set***. Now use the training set to select and fit a model. Then see how well the model fitted to the training set fits the validation set.

- In either case, a measure of how well the model fitted to the original (or training) data set fits the new (or validation) set is the ***mean squared prediction error***, denoted **MSPR**:

Mean Squared Prediction Error:

$$\text{MSPR} \;=\; \frac{1}{n*}\sum_{i=1}^{n*}(Y_i - \hat{Y}_i)^2$$

where

$$\begin{aligned} n* &= \text{ The sample size for the new (or validation) data set.} \\ Y_i &= \text{ The } i\text{th observation in the new (or validation) set.} \\ \hat{Y}_i &= \text{ The predicted value for } Y_i \text{ based on the model that was} \\ &\quad\ \text{selected and fitted to the original (or training) data set.} \end{aligned}$$

- The **MSPR** should be compared to the **MSE** for the original (or training) data set:

  ▷ A **MSPR** fairly **close** to the **MSE** indicates that the **model** is **valid**.

  ▷ A **MSPR** much **larger** than the **MSE** suggests that the **model *overfits*** the original (or training) data set.

- ***Overfitting*** results when the ***model complexity*** is too high. In regression, **model complexity** is determined by the **number** of **terms** in the model.

- For example, for the data shown in Fig. 1, we can fit **two models**:

$$\begin{aligned} \text{Model 1}: \quad Y &= \beta_0 + \beta_1 X + \epsilon \\ \text{Model 2}: \quad Y &= \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \epsilon \end{aligned}$$

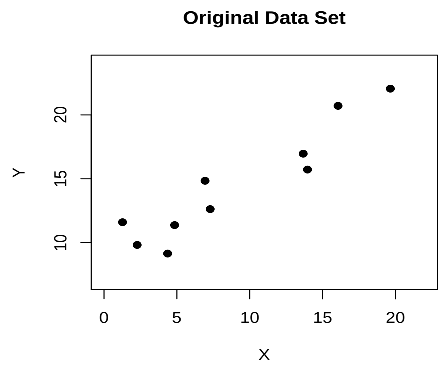
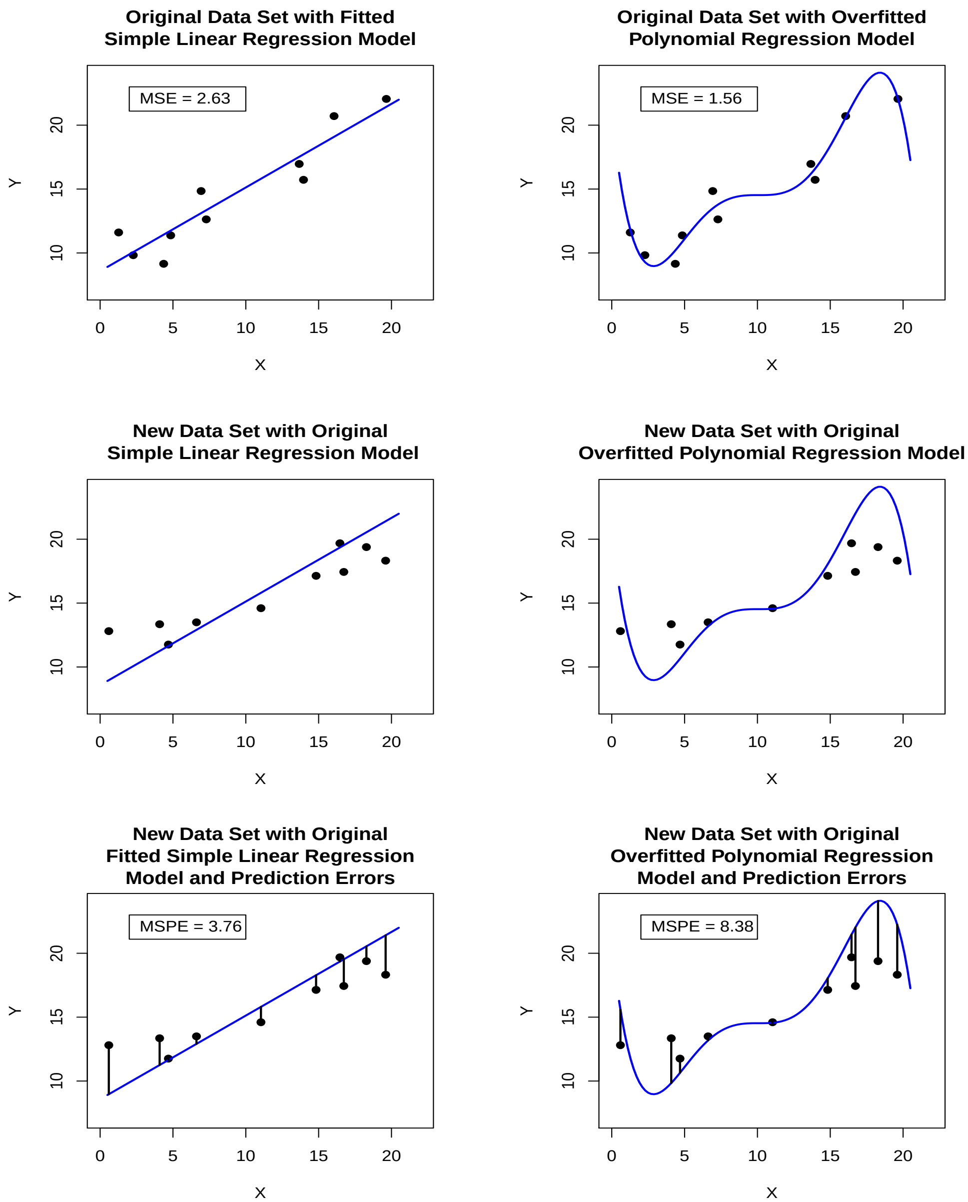


Figure 1

The fitted models are shown below with the **original** data and **new** data.



2 Model Diagnostics

- Model adequacy can be further assessed using various **diagnostic measures**.

2.1 Identifying Outlying Y Observations

- To identify **outlying $\boldsymbol{Y}$ observations**, it's useful to **standardize** the **residuals** using an **estimated standard deviation**.

  There are two ways to estimate the **standard deviation** of the **residuals**.

  Depending on which way is used, the standardized residuals are called either ***semistudentized*** or ***studentized*** residuals.

2.1.1 Semistudentized Residuals

- Recall that the ith residual is $e_i = Y_i - \hat{Y}_i$. The ith ***semistudentized residual***, denoted $\boldsymbol{e_i^*}$, is

  Semistudentized Residuals:

  $$e_i^* \;=\; \frac{e_i}{\sqrt{\text{MSE}}} \tag{1}$$

  Values of e_i^* **larger** (in absolute value) than about **3** or **4** should be investigated as potential **outliers** in the Y variable.

2.1.2 Studentized Residuals

- We know that MSE is an estimator for σ^2, the true variance of the $\text{N}(0, \sigma^2)$ **error term $\boldsymbol{\epsilon}$** in the regression model.

  But the true variance of the **$\boldsymbol{i}$th residual $\boldsymbol{e_i}$**, denoted $\boldsymbol{\sigma^2\{e_i\}}$, is *not* σ^2.

  Rather, it can be shown that

  $$\sigma^2\{e_i\} \;=\; \sigma^2 \cdot (1 - h_{ii})\,, \tag{2}$$

  where $\boldsymbol{h_{ii}}$ is the ith diagonal element of the $n \times n$ ***hat matrix*** $\mathbf{H}$ (Class Notes 11), defined as

  $$\mathbf{H} \;=\; \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,.$$

- Note that because the diagonal elements h_{ii} of $\mathbf{H}$ generally are not all equal, the residuals have unequal variances $\sigma^2\{e_1\}, \sigma^2\{e_2\}, \ldots, \sigma^2\{e_n\}$.
- An estimator of $\sigma^2\{e_i\}$ is $\mathbf{MSE} \cdot (\mathbf{1} - \boldsymbol{h_{ii}})$, and the ith ***studentized residual***, denoted $\boldsymbol{r_i}$, is defined as

Studentized Residuals:

$$r_i = \frac{e_i}{\sqrt{\text{MSE} \cdot (1 - h_{ii})}} \qquad (3)$$

Values of r_i **larger** (in absolute value) than about **3** or **4** should be investigated as potential **outliers** in the Y variable.

2.1.3 (Optional Section) Deleted Residuals and Studentized Deleted Residuals

- An outlier Y_i in the Y variable can influence the fitted regression line (or hyperplane), making that outlier less noticeable. It's sometimes useful, then, to fit the model to the data with the ith observation omitted, and then calculate the residual corresponding to the deviation of Y_i away from the model just fitted.

  The ith ***deleted residual*** $\boldsymbol{d_i}$ is defined as

  Deleted Residuals:

  $$d_i = Y_i - \hat{Y}_{i(i)}$$

  where $\hat{Y}_{i(i)}$ is the fitted value for the ith individual based on the model fitted to the data with ith observation omitted.

  These residuals were used to compute **PRESS** (Class Notes 19).

- It can be shown that an estimator for the variance $\sigma^2\{d_i\}$ of d_i is

  $$s^2\{d_i\} = \frac{\text{MSE}_{(i)}}{1 - h_{ii}},$$

  where $\text{MSE}_{(i)}$ is the mean squared error obtained after fitting the model with the ith observation left out.

- The ith ***studentized deleted residual*** $\boldsymbol{t_i}$ is defined as

  Studentized Deleted Residual:

  $$t_i = \frac{d_i}{\sqrt{\text{MSE}_{(i)}/(1 - h_{ii})}} = \frac{e_i}{\sqrt{\text{MSE}_{(i)} \cdot (1 - h_{ii})}}$$

Values of t_i **larger** (in absolute value) than about **3** or **4** should be investigated as potential **outliers** in the Y variable.

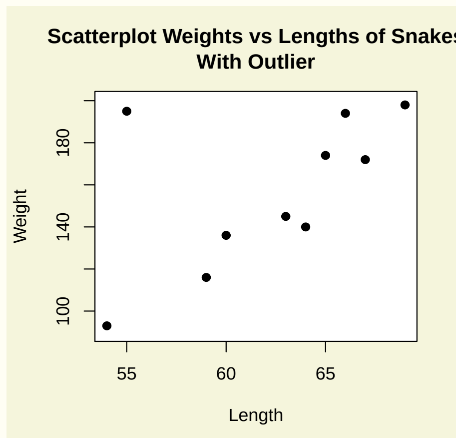
2.2 Identifying Outlying X Observations

- Outliers among the ***multivariate*** **predictor values** X can be especially **influential** on the fitted regression line or plane (or hyperplane).
- ***Multivariate outliers*** can be **difficult** to **detect** in **two-dimensional scatter-plots**, where only two variables can be plotted at a time.

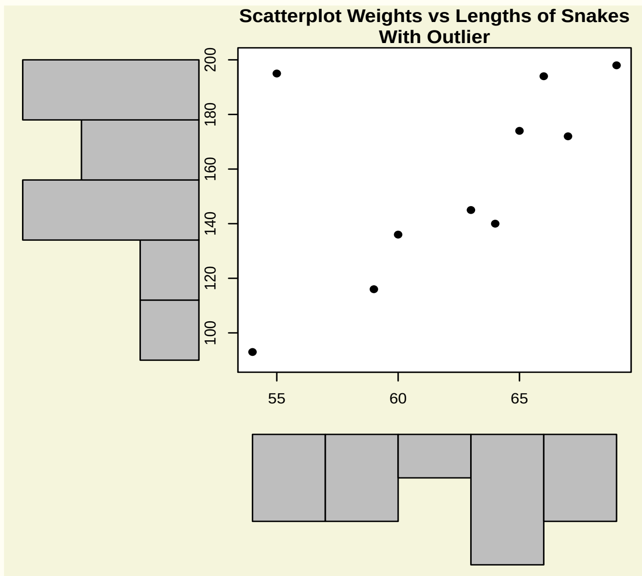
Example 2.1 An **outlying tenth snake** has been added to the data and scatterplot below on lengths and weights of female snakes (Class Notes 1).

Lengths and Weights of Female Snakes With Outlier

Snake	Length (cm)	Weight (g)
1	60	136
2	69	198
3	66	194
4	64	140
5	54	93
6	67	172
7	59	116
8	65	174
9	63	145
10	**55**	**195**



The **two-dimensional outlier** doesn't show up in either of the **one-dimensional plots** (histograms) shown below.



- So we'll need a metric for identifying outlying multivariate X observations.

  One measure of whether the ith (multivariate) X observation X_i is an outlier is the ***leverage***, defined as the **ith diagonal element h_{ii}** of the **hat matrix H**.

  It can be shown that a (multivariate) observation $X_i = (X_{i1}, X_{i2}, \ldots, X_{i,p-1})$ that's **far** from the (multivariate) ***centroid*** $(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_{p-1})$ of the X observations will have a **large h_{ii}** value.

  Such **outlying X** observations tend to have **more influence** on the fitted regression line or plane (or hyperplane).

- The following helps explain why h_{ii} measures the **influence** of the ith observation:

  1. $\hat{\mathbf{Y}} = \mathbf{HY}$, where $\mathbf{H}$ is the hat matrix, $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$, and $\hat{\mathbf{Y}} = (\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n)^T$, so the ith fitted value is

  $$\hat{Y}_i \;=\; h_{i1}Y_1 + h_{i2}Y_2 + \cdots + h_{ii}Y_i + \cdots + h_{in}Y_n\,, \tag{4}$$

     where h_{ij} is the (i,j)th element of $\mathbf{H}$.
  2. We see from (4) that the leverage h_{ii} measures the contribution (influence) of the ith individual's response Y_i in determining the ith fitted value $\hat{Y}_i$.

- It can be shown that $0 \leq h_{ii} \leq 1$.

  Values of h_{ii} **larger** than **0.5** indicate **highly influential** observations. Values **between 0.2** and **0.5** indicate **moderately influential** observations.

2.3 Identifying Influential Observations

After identified outlying Y or (multivariate) X observations, we'll need a few measures of their **influence** on the fitted model. We'll look at:

1. DFFITS
2. Cooke's Distance
3. DFBETAS

2.3.1 DFFITS

- One way to deem an observation **influential** is to show that its deletion from the data set would result in a dramatic change in the regression **fitted values**.
- The ith value of ***DFFITS*** ("difference in fitted values") is defined to be

> **DFFITS**:
> $$(\text{DFFITS})_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_{ii}}},$$
> where $\hat{Y}_i$ and $\hat{Y}_{i(i)}$ are the fitted values for the ith individual based on models fitted, respectively, to the complete data set and the data set with ith observation omitted, $\text{MSE}_{(i)}$ is the mean squared error obtained after fitting the model with the ith observation omitted, and h_{ii} is the ith diagonal element of the hat matrix.

There will be n different $(\text{DFFITS})_i$ values, one for each individual in the data set.

$(\text{DFFITS})_i$ represents the (standardized) **change** in the ***i*th fitted value** that would result if the ***i*th observation** was **omitted** from the data set.

Values of $(\text{DFFITS})_i$ **larger** (in absolute value) than **1** indicate **influential** observations.

2.3.2 Cooke's Distance

- ***Cooke's distance*** is another measure the effect of removing the ith observation from the data on the fitted line or hyperplane.
- The ith ***Cooke's distance*** value $\boldsymbol{D_i}$ is defined as

Cooke's Distance:

$$D_i = \sum_{j=1}^{n} \frac{(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\,\text{MSE}},$$

where $\hat{Y}_i$ and $\hat{Y}_{i(i)}$ are the fitted values for the ith individual based on models fitted, respectively, to the complete data set and the data set with ith observation omitted, p is the number of parameters in the model, and MSE is the mean squared error after fitting the model to the full data set.

There will be n different D_i values, one for each individual in the data set.

Cooke's distance D_i measures the influence of the ith observation on **all** the fitted values, unlike DFFITS, which only measures its influence on the ith fitted value.

A **large** D_i value suggests that the ith observation is **influential**. Values **larger** than the **50th percentile** (median) of the $\boldsymbol{F(p, n-p)}$ **distribution** indicate **influential** observations.

2.3.3 DFBETAS

- Another way to deem an observation **influential** is to show that its deletion from the data set would result in a dramatic change in the **estimated regression coefficients**.
- For each each estimated coefficient b_k, the influence of the ith observation on b_k is measured by the ith ***DFBETAS*** value ("difference in estimated betas") for that coefficient, defined as

DFBETAS:

$$(\text{DFBETAS})_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{\text{MSE}_{(i)} c_{kk}}}$$

where b_k and $b_{k(i)}$ are the estimates of β_k using, respectively, the complete data set and the data set with ith observation omitted, $\text{MSE}_{(i)}$ is the mean squared error obtained after fitting the model with the ith observation omitted, and

c_{kk} is the kth diagonal element of the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$.

For each of the coefficients $b_0, b_2, \ldots, b_{p-1}$, there will be n $(\text{DFBETAS})_{k(i)}$ values, one for each individual in the data set.

$(\text{DFBETAS})_{k(i)}$ represents the (standardized) **change** in the $\boldsymbol{k}$**th estimated coefficient** $\boldsymbol{b_k}$ that would result if the $\boldsymbol{i}$**th observation** was **omitted** from the data set.

A **large** $(\text{DFBETAS})_{k(i)}$ value indicates that the ith observation is **influential** on the kth estimated coefficient. For small to medium sized data sets, values **larger** than **1** indicate **influential** observations, and for large data sets, values **larger** than $\boldsymbol{2/\sqrt{n}}$ do.

2.4 Detecting Multicollinearity: The Variance Inflation Factor

- Recall the **problems** that arise from **multicollinearity** include:

  1. The coefficient estimates and their p-values and t test statistic values can change depending on which other predictors are included in the model.
  2. Extra sums of squares and partial F test results depend on the order in which predictors are added into the model.
  3. Standard errors of estimated coefficients can be very large.
  4. The regression model F test result may be statistically significant even though none of the t test results for individual coefficients are significant.

  Any of the above conditions can be used to **detect** multicollinearity informally.

- A more formal way to **detect** (and **measure**) **multicollinearity** is the kth ***variance inflation factor***, denoted $\boldsymbol{(\text{VIF})_k}$ and defined as

Variance Inflation Factor:

$$(\text{VIF})_k \;=\; \frac{1}{1-R_k^2}$$

where R_k^2 is the coefficient of multiple determination when X_k is regressed (as the response variable) on the other $p-2$ predictors.

There will be one $(\text{VIF})_k$ value for each of the $p-1$ predictors in the model.

It can be shown that $(\text{VIF})_k$ is an indicator of how much the estimated **variance** (squared standard error) of $\boldsymbol{b_k}$ is **"inflated"** as a result of the correlations between X_k and the other predictors. See the textbook.

If $R_k^2 = 0$ (i.e. X_k is not related to the other predictors), then $(\text{VIF})_k = 1$, meaning that multicollinearity does not inflate the variance of b_k. But if $R_k^2 > 0$, then $(\text{VIF})_k > 1$, meaning that the variance of b_k is inflated as result of multicollinearity.

A **large** $(\text{VIF})_k$ is a sign of **multicollinearity**. If the *largest* of the $(\text{VIF})_k$ values is **greater** than **10**, it indicates that multicollinearity may be a problem and that the estimated coefficients, t tests, etc. may be unreliable.